

Filterless Snapshot Hyperspectral Imaging using Guided Patch Diffusion

Dean Hazineh[✉], Luca Sacchi, Davide Cassara[✉],
Federico Capasso[✉], and Todd Zickler[✉]

Harvard University
dhazineh@g.harvard.edu

Abstract. We consider the problem of reconstructing a $H \times W \times 31$ hyperspectral image from a $H \times W$ grayscale snapshot measurement that is captured using only a single diffractive lens and a filterless panchromatic photosensor. This problem is severely ill-posed, but we present a model that produces high-quality results in simulation and experiment. We make efficient use of limited training data by creating a conditional denoising diffusion model that operates on small patches in a shift-invariant manner. During inference, we synchronize per-patch hyperspectral predictions using guidance by physical consistency with the system’s optical point spread function. Our experiments reveal that the patch size can be as small as the point spread function, with local optical cues being the main source of information about complete spectra. Also, by drawing multiple samples, our model provides per-pixel uncertainty estimates that strongly correlate with reconstruction error.

Keywords: Computational Imaging · Hyperspectral Imaging · Diffusion Models

1 Introduction

Snapshot hyperspectral cameras capture detailed spectral information about a scene at a single moment in time. They offer a richer representation than standard RGB images and are useful for scientific detection and classification. Generally, these cameras have two coupled components: an optical assembly that encodes spatial and spectral information onto a photosensor, and a digital decoder that reconstructs the hyperspectral image (HSI) from the resulting measurement. To better condition the reconstruction problem, existing snapshot designs typically use one or more of the following strategies [12]: complex, multi-stage optics; color filter arrays on the photosensor; and/or photosensors with more pixels than the intended HSI spatial resolution.

In this paper, we explore a minimalist snapshot scenario that is less well-posed and largely underexplored. Our goal is to reconstruct a $H \times W \times 31$ HSI using only: (*i*) a filterless panchromatic photosensor with $H \times W$ pixels, the same

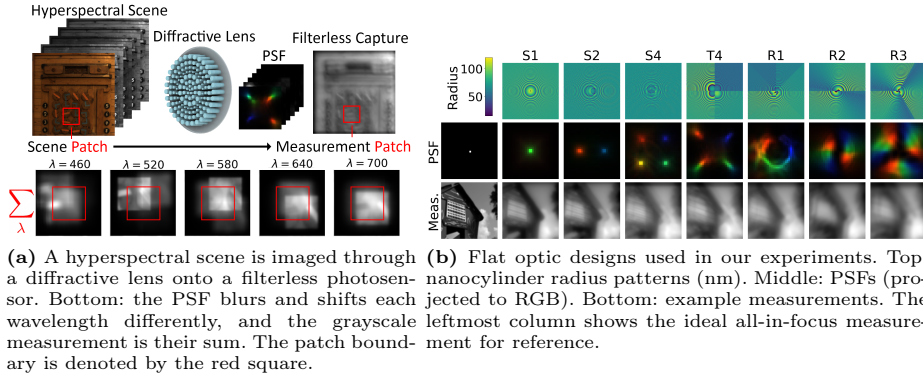


Fig. 1: Forward optical model and flat optic designs.

number of measurement pixels as output pixels; and (ii) a single flat optic lens, such as a diffractive optical element or a metalens. This scenario is interesting because it could enable a new class of snapshot hyperspectral cameras with distinct trade-offs between optical and computational complexity. Using a single flat optic improves compactness, and removing spectral filter mosaics on the sensor increases light throughput and avoids interleaved sub-sampling.

We can design the flat optic to induce purposeful chromatic aberration that mixes spatial and spectral information into the filterless measurement, as shown in the top of Fig. 1a. However, reconstructing the HSI from this mixture is severely ill-posed. It requires an effective prior from a deep learning model, but this is hard to train because, compared to RGB images, ground-truth hyperspectral images are relatively scarce. Patch-based generative diffusion models have recently emerged as a promising solution for learning strong priors from small datasets [15, 20, 21, 43, 52], but conventional patch-based processing is particularly difficult to apply here. As shown in the bottom of Fig. 1a, the measurement is formed by convolution with an optical point-spread function (PSF) with a spatial extent that is comparable to the patch size. This means that a significant amount of the target hyperspectral signal is scattered outside of its corresponding measurement patch (red square), making per-patch reconstruction very ambiguous. As far as we know, no prior patch-based approach to reconstruction, diffusion-based or otherwise, has been shown to have success in these conditions.

We address this challenge and introduce a patch diffusion model that succeeds at our task. We overcome the ambiguity associated with patching by adopting global diffusion guidance during inference, where patches are iteratively denoised in parallel and then assembled into full-sized HSIs that are forced to be optically consistent with the full-size measurement. We find that this resolves patch-based ambiguities and provides better results than any previous model applied to our task. Like any patch-based approach, our model has the advantage of being able to operate on any image size once it is trained, and we also find that it provides useful per-pixel uncertainty estimates for its spectral predictions that

strongly correlate with reconstruction error. We extensively evaluate our method in simulation and provide an experimental prototype as a proof-of-concept.

2 Related Work

Training diffusion models on image patches substantially reduces data requirements. Unconditional patch models can generate high-quality RGB images from as few as 5K samples [43], and patch-based diffusion has been applied to inverse problems including deblurring [21] and deweathering [52] using comparably small datasets. However, these works consider spatially compact degradations. We show that patch-based diffusion can handle very large degradation kernels by leveraging inference-time guidance.

In aerial remote sensing, where large hyperspectral datasets exist, unconditional diffusion models have been trained from scratch to learn representations for classification [8, 38]. For natural scenes, datasets are far smaller—ARAD1K [2], for example, contains only 950 HSIs—and existing methods instead rely on frozen, pre-trained RGB diffusion models for plug-and-play HSI restoration [35, 49] or compressed sensing [34]. In contrast, we train conditional hyperspectral diffusion models from scratch on natural scene patches, learning spatial-spectral patch priors without overfitting to the small available datasets. To our knowledge, similar models have not been explored in this context.

Recovering a HSI from a single unfiltered measurement has traditionally required complex optical assemblies. CASSI [31, 41], for example, uses coded masks, dispersive elements, and oversized photosensors (*e.g.*, $H \times (W+30)$ measurement pixels for an $H \times W \times 31$ output). Other approaches rely on spectral filters such as RGB Bayer patterns [1, 2, 5, 50] (sometimes paired with diffractive lenses [24, 30, 51]) or optimized color filter arrays [27, 32, 33, 37]. However, spectral filters discard a large fraction of incoming light—a Bayer mosaic, for example, captures only one of three color channels per pixel. Our approach requires neither multi-component optics nor spectral filters: we place a single flat diffractive lens in front of a bare panchromatic photosensor and rely on a strong diffusion-based decoder to recover the HSI. Concurrently, Liu *et al.* [29] explore a similar, filterless reconstruction problem but using a SLM-based optical encoder, providing further evidence that this problem class is tractable.

Additionally, we leverage our diffusion model to produce uncertainty estimates for predicted spectra. Concurrent work from Romero *et al.* [36] also explores uncertainty quantification in HSI reconstructions using diffusion posterior sampling, though applied to RGB inputs.

3 Methods

Let $\mathbf{x} \in \mathbb{R}_{\geq 0}^{H \times W \times C}$ denote a hyperspectral image (HSI), representing the far-field scene’s undistorted spatial-spectral radiance mapped to the photosensor plane by an ideal lens focused at infinity. This representation accounts for geometric magnification and spatial discretization to the sensor’s pixel size. We define the

associated measurement $\mathbf{y} \in \mathbb{R}_{\geq 0}^{H \times W}$ that is induced by a diffractive lens using the element’s wavelength-dependent point-spread function (PSF) $f(u, v, \lambda)$ as

$$\mathbf{y}(u, v) = \mathcal{M}(\mathbf{x}) = \sum_{\lambda} o(\lambda) \cdot f(u, v, \lambda) \underset{(u,v)}{*} \mathbf{x}(u, v, \lambda), \quad (1)$$

where $*$ denotes 2D convolution over the spatial dimensions and $o(\lambda)$ corresponds to the spectral response of the photosensor. Throughout, we assume that the PSF is shift-invariant. A measurement is thus a non-invertible linear optical encoding of a 3D hyperspectral cube to a 2D image. Our objective is to model the conditional distribution $p(\mathbf{x} | \mathbf{y})$ and to sample hyperspectral images $\mathbf{x}$ that are consistent with the observed measurement $\mathbf{y}$. We do this by training a patch-based diffusion model and introducing a guided sampling algorithm that synchronizes patch predictions into measurement-consistent full-field HSIs.

3.1 Optical Encoding

In our experiments, we test the eight point-spread functions (PSFs) shown in the middle row of Fig. 1b to understand what type of optical encoder is most effective for our task. These PSFs vary in the extent to which they spread spectral information across space, producing differently-blurred measurements. More concentrated PSFs (left) produce sharper images that preserve spatial detail but encode spectral information less effectively than more dispersive PSFs (right). Because reconstruction requires both high spatial and spectral accuracy, it is not obvious which type of PSF will perform the best in our filterless scenario.

All of these PSFs can be physically realized using a diffractive lens known as a *metalens*—a transparent glass sheet patterned with nanoscale cylinders of equal height and varying widths [25, 26]. The radius of each nanocylinder controls the local phase-delay, and each of the PSFs results from a different arrangement of radii (top). We design a subset of the lenses, labeled with prefix “S”, using spatial multiplexing to produce a quasi-stationary, multi-foci effect [3]. The other lenses, labeled with prefix “T” or “R”, are designed using angular multiplexing to produce PSFs whose focal spots shift or rotate as a function of wavelength. The designs for “R2” and “R3” have been used previously for RGB-to-hyperspectral imaging and follow from [24]. The PSF for each lens is computed using a wave-optics simulator [4, 16, 17]. See Supp. Sec. S1 for more details.

3.2 Denoising Diffusion on Patches

We use a conditional denoising diffusion model [18] to sample from $p(\mathbf{x} | \mathbf{y})$. The forward process corrupts a clean HSI $\mathbf{x}_0$ with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ according to $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$, where α_t is a variance schedule. A network $\epsilon_{\theta}(\mathbf{x}_t, t; \mathbf{y})$ is trained to predict ϵ from the noisy sample $\mathbf{x}_t$ at timestep t , conditioned on the measurement $\mathbf{y}$, by minimizing $\mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t; \mathbf{y})\|^2]$. Samples

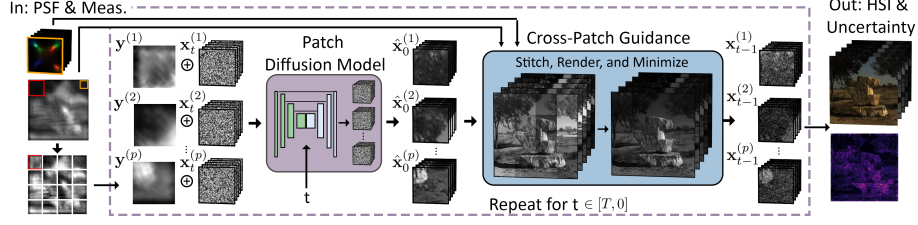


Fig. 2: Overview of the guided sampling pipeline. The input measurement is split into patches, each concatenated with a noise sample and denoised. Intermediate patch predictions are stitched into a full-field HSI and passed through the forward model to compute a guidance gradient, which updates all patches before the next reverse diffusion step. Multiple runs with different random seeds yield per-pixel uncertainty.

are drawn via the reverse process [40]:

$$\begin{aligned}\hat{\mathbf{x}}_0(\mathbf{x}_t) &= \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_\theta}{\sqrt{\alpha_t}}, \\ \mathbf{x}_{t-1} &= \sqrt{\alpha_{t-1}} \hat{\mathbf{x}}_0(\mathbf{x}_t) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \boldsymbol{\epsilon}_\theta + \mathbf{w},\end{aligned}\tag{2}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \mathbb{I})$ and σ_t controls stochasticity. The intermediate estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ plays a central role in our guided sampling procedure (Sec. 3.3).

Rather than denoising full HSIs at once, we model the conditional distribution at the patch level. Because the PSF support is comparable to the patch size, each measurement patch contains signal from neighboring HSI regions while also losing signal scattered to adjacent patches (Fig. 1a). Despite this crosstalk, we find that a patch-based diffusion model can learn a prior that compensates for the ambiguous information at patch boundaries. For training data, we use captured HSIs from a ground truth hyperspectral camera and pre-render the corresponding measurements via Eq. (1). We then train our models using pairs of patches $(\mathbf{x}_0^{(i)}, \mathbf{y}^{(i)})$ randomly cropped from the full-field HSI-measurement pairs. We implement conditioning through concatenation, as shown in Fig. 2. Each patch pair is max-normalized prior to training (see Supp. Sec. S3.2); the resulting per-patch scale ambiguity is resolved during guided sampling as described next.

3.3 Sampling with PSF Guidance

Applying the denoising formulation in Eq. (2) to patches produces hyperspectral patch predictions $\hat{\mathbf{x}}_0^{(i)}$ at intermediate time steps t . We use these predictions to guide the denoising step and impose additional constraints when sampling $\mathbf{x}_{t-1}^{(i)}$ from $\mathbf{x}_t^{(i)}$ [10, 11]. In particular, we enforce that all intermediate hyperspectral patches stitch together into a full-field HSI that is optically consistent with the full-field measurement. Pseudo-code is provided in Supp. Sec. S3.3; here

we summarize the key steps. Throughout, we use superscript p to denote a p -element collection of patches, *e.g.* $\mathbf{x}_t^p = \{\mathbf{x}_t^{(i)}\}_{i=1}^p$ and define a $\text{Stitch}(\cdot)$ operator that combines those patch estimates into a single full-field HSI. The operator $\mathcal{M}(\cdot)$ refers to the measurement operation in Eq. (1).

During deployment, we split the full-field measurement $\mathbf{y}$ into non-overlapping patches $\mathbf{y}^p$, each concatenated with a per-patch noise sample $\mathbf{x}_T^p$. We then process these patches in parallel through a denoising step to obtain the intermediate denoised estimates $\hat{\mathbf{x}}_0^p$. Next, we stitch those estimates into a full-field HSI and pass it through the measurement operator. Because each patch is max-normalized during training, the denoised estimates are accurate up to an unknown per-patch scale. We recover these scales by solving the least-squares problem,

$$c_{\text{lsq}}^p = \underset{c^p}{\operatorname{argmin}} \|\mathcal{M}(\text{Stitch}(c^p \cdot \hat{\mathbf{x}}_0^p(\mathbf{x}_t^p))) - \mathbf{y}\|^2, \quad (3)$$

which admits a closed-form solution and is computed in a single pass. With the scale factors fixed at c_{lsq}^p , we define a guidance loss over the patch states to measure consistency with the full-field measurement,

$$\mathcal{L}(\mathbf{x}_t^p, \mathbf{y}) = \|\mathcal{M}(\text{Stitch}(c_{\text{lsq}}^p \cdot \hat{\mathbf{x}}_0^p(\mathbf{x}_t^p))) - \mathbf{y}\|^2. \quad (4)$$

This loss provides a gradient-based correction to all patch states, applied via the modified denoising transition,

$$\tilde{\mathbf{x}}_t^p = \mathbf{x}_t^p - \eta \nabla_{\mathbf{x}_t^p} \mathcal{L}(\mathbf{x}_t^p, \mathbf{y}) / \|\nabla_{\mathbf{x}_t^p} \mathcal{L}(\mathbf{x}_t^p, \mathbf{y})\| \quad (5)$$

$$\mathbf{x}_{t-1}^p = \sqrt{\alpha_{t-1}} \hat{\mathbf{x}}_0^p(\tilde{\mathbf{x}}_t^p) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \epsilon_\theta^p + w^p, \quad (6)$$

where $\tilde{\mathbf{x}}_t^p$ and $\mathbf{x}_{t-1}^p$ are the updated patch states. We repeat the gradient step in Eq. (5) multiple times before the denoising step in Eq. (6). Fig. 2 provides an overview of the entire sampling pipeline.

Finally, by varying the initial noise samples $\mathbf{x}_T^p$, we obtain multiple plausible HSIs from the same measurement. We quantify spectral uncertainty by computing a per-pixel variance map from N draws,

$$\text{Uncertainty} = \sum_{\lambda} \text{Var}(\{\mathbf{x}_0\}_{i=1}^N). \quad (7)$$

4 Simulation Results

We first evaluate our reconstruction algorithm in simulation. We use the ARAD1K dataset [2] with a standard 900/50 train/test split, reconstructing 31 spectral channels uniformly spanning 400–700 nm. Unless otherwise stated, we resize HSIs to 256×256 pixels, matching the resolution used by competing approaches; these methods process full-field measurements through a single network and cannot easily scale higher. Our patch-based model does not share this limitation. Grayscale measurements are rendered using the R1 PSF, which yielded the best

Table 1: Filterless reconstruction on ARAD1K using the R1 PSF. Metrics: (S)AM, (SS)IM, (P)SNR.

Model	S ↓	SS ↑	P ↑
Ours	0.11	0.94	34.6
Ours (no guid.)	<u>0.14</u>	<u>0.92</u>	<u>32.32</u>
UNet-Regress.	0.15	0.83	29.12
SST [28]	0.15	0.90	31.8
SPECAT [44]	0.18	0.84	29.6
MST [5]	0.17	0.87	29.8
In2Set [42]	0.18	0.86	30.1
DAUHST [6]	0.17	0.86	29.7
DGSMP [23]	0.16	0.88	30.0
HDNet [22]	0.17	0.86	29.3
TSANet [31]	0.20	0.87	29.2

Table 2: Model ablations. [†]uses overlapping patches with a stride of 32 pixels

Patch	Resc/Guid	SS ↑	P ↑
64	✓/✓	0.94	34.7
64	✓/✗	0.92	32.2
64 [†]	✓/✓	0.95	34.8
64 [†]	✓/✗	0.93	33.0
32	✓/✓	0.93	33.3
32	✓/✗	0.87	29.3
64	✗/✓	0.92	31.8
64	✗/✗	0.86	27.5

Table 3: Lens comparison.

*designs from [24]

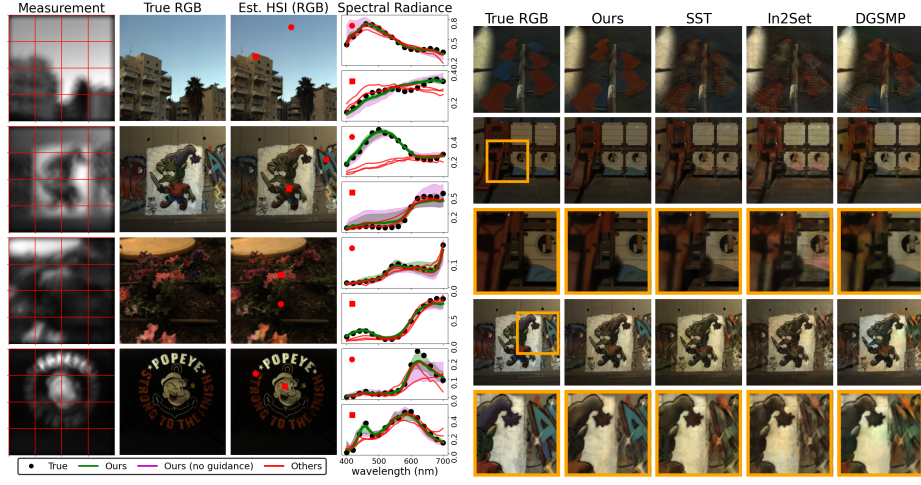
Lens	S ↓	SS ↑	P ↑
AIF	0.20	0.93	29.8
S1	0.17	0.93	31.1
S2	0.15	0.94	33.1
S4	0.13	0.94	34.4
T4	<u>0.11</u>	<u>0.94</u>	<u>34.6</u>
R1	0.11	0.95	35.0
R2*	0.11	0.93	34.0
R3*	0.12	0.93	34.1

empirical performance (Sec. 4.4). We train on 64×64 patches and sample using 20 DDIM steps and 20 guidance iterations (Sec. 4.3). We report SSIM, PSNR, and SAM [47] averaged over full-field reconstructions. Additional architecture, training, and inference details are provided in Supp. Sec. S3.

4.1 Comparison to Other Models

We compare our patch-based diffusion approach to eight state-of-the-art hyperspectral reconstruction models [5,6,22,23,28,31,42,44] that map fixed-resolution, full-field measurements to full-field HSIs. As no prior models exist for our single-optic filterless setting, these represent the closest available baselines from related but better-conditioned snapshot HSI tasks. We train all models from scratch on our rendered measurements following their original training procedures, making only minimal modifications where necessary, *e.g.* replacing their forward/adjoint operators with our measurement function and changing the number of output channels. In Supp. Sec. S4.2, we repeat this comparison using RGB Bayer-filtered measurements, where our method also achieves the best performance.

The filterless results are shown in Tab. 1 and visualized in Fig. 3, with additional visualizations in Supp. Sec. 4.5. Our model achieves an average PSNR of 34.63 dB, surpassing the next best method by 2.86 dB. We also obtain a higher SSIM (0.94 vs. 0.90) and a lower SAM (0.11 vs. 0.15), reflecting more accurate per-pixel spectral predictions. For comparison, we also train our model’s UNet backbone as a single-shot reconstruction network using a standard MSE regression loss on the full HSI instead of patches (“UNet-Regress” in Tab. 1). This variant under-performs all models, suggesting that our gains come not from increased backbone capacity but instead from our patch-based diffusion. These results also suggest that focusing model capacity on local optical cues in patches while enforcing global consistency through guidance is more effective than processing the entire measurement with a larger receptive field.



(a) Left columns: grayscale measurement, ground truth RGB, and our estimated HSIs displayed in RGB. Right column: spectral profiles at two marked pixels per scene. Green: our mean reconstructed estimate with uncertainty (fill). Magenta: without guidance. Red: three next-best baselines. Black dots: ground truth. (b) Comparison of reconstructions (projected to RGB) from our model and leading baselines for three test scenes. Zoomed-in regions (orange squares) are shown in the third and fifth rows.

Fig. 3: Simulated grayscale-to-hyperspectral reconstructions on the ARAD1K test set.

In contrast to other models, our method also provides pixel-wise uncertainty estimates. These uncertainty maps, computed via Eq. (7), are shown for several test scenes in Fig. 4. We find that the per-pixel uncertainty is strongly correlated with the mean squared error (MSE) between the predicted and ground-truth HSIs. Specifically, we compute a Pearson correlation coefficient of 0.80 using 12.5K randomly sampled pixels from the 50 test images. This correlation remains strong when analyzed per-wavelength channel (Supp. Sec. S4.1). These results suggest that the uncertainty estimates may be useful in practice for identifying regions where the reconstruction is less reliable.

4.2 Cross-Dataset Generalization

We apply our diffusion model to reconstruct HSIs from other datasets, some at different spatial resolutions, without further finetuning. We train on ARAD1K at its native resolution of 512×512 pixels and then reconstruct measurements rendered from ICVL [1] (1280×1280 ; 400 patches), Harvard [7] (1024×1344 ; 336 patches), and CAVE [45] (512×512 , 64 patches). Figure 5 displays the reconstruction results for selected scenes, with additional visualizations in the supplement. Averaging over 10 test scenes, we obtain a mean PSNR/SSIM of 33.48/0.94 for ICVL and 32.37/0.92 for Harvard, indicating that our model generalizes well when the optical model is unchanged. We also observe wider uncertainty spreads in the predicted spectra (magenta and green fill in Fig. 5

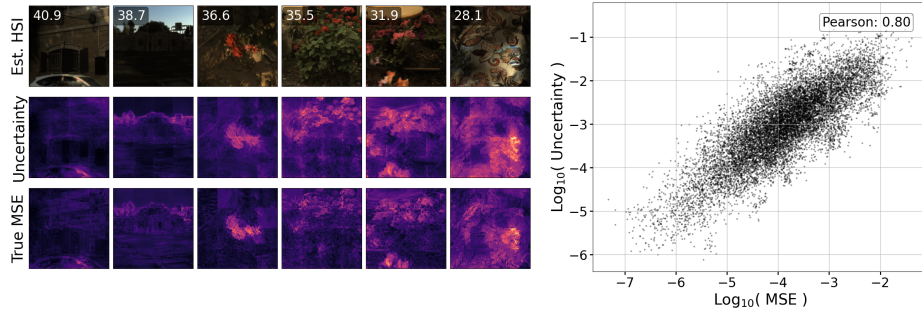


Fig. 4: (Left) Estimated HSIs projected to RGB with per-pixel uncertainty maps and true MSE. Images are overlaid with PSNR and arranged from highest on the left to lowest on the right. (Right) Scatter plot of estimated uncertainty vs. true error from sampled test pixels.

as compared to in Fig. 3a), reflecting the model’s increased uncertainty under distribution shift. In the rightmost column of Fig. 5, we highlight our model’s reconstruction when applied to an RGB-metameric scene from the CAVE dataset. Following [13], we reconstruct a scene containing real and fake fruits, where pairs of objects have spectral radiances that differs but RGB tristimulus values that are nearly identical. We find that our model can directly discriminate between metameric pixels without finetuning. We attribute this property to the fact that our reconstructions are driven by local optical cues (Sec. 4.4) rather than learned object-level semantics.

4.3 Run Time and Model Ablations

We apply guidance by taking a gradient descent step on $\mathbf{x}_t$ (Eq. (5)) that is *regularized* by the diffusion model, repeating it for several iterations before the next denoising step $\mathbf{x}_{t-1}$ (Eq. (6)). Figure 6a shows that reconstruction accuracy improves logarithmically with the number of guidance iterations, with PSNR increasing from 32.32 (no guidance) to 35.22 at 60 iterations. Patches are processed in parallel, so complexity is dominated by the convolution in the measurement operator, which scales as $O(n \log n)$ for n patches via FFTs. This guidance step is the main computational bottleneck; we use gradient checkpointing to reduce memory at the cost of longer run times [9]. Reconstruction requires approximately 22 seconds per scene for ARAD1K, but up to 52 minutes per scene for the higher-resolution Harvard and ICVL on an RTX 3090 GPU. Our method is thus suited to applications that require fast capture but not fast reconstruction. This cost can be reduced in several ways. We empirically find that guidance is more beneficial in later denoising steps, so the iteration count can be scheduled across steps rather than held fixed. For example, increasing the iteration count via a quadratic schedule over the DDIM steps reduces reconstruction time by $3\times$ on the ARAD1K test set without a loss in quality. Our method is also compat-

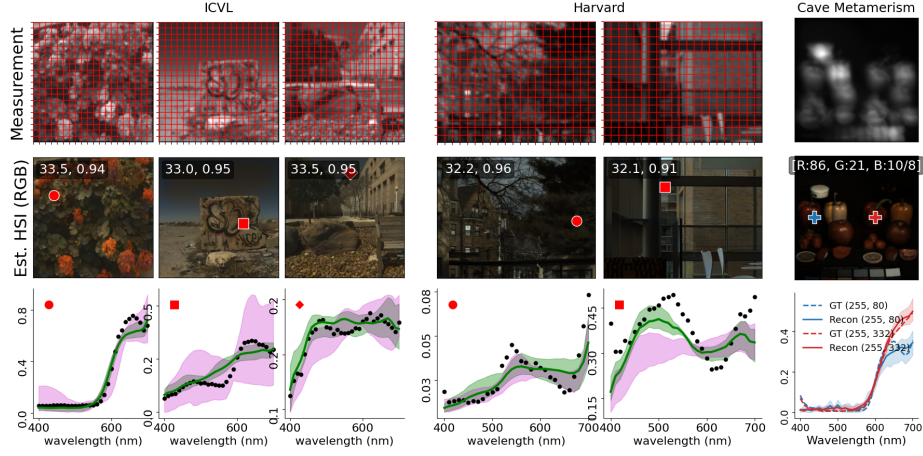


Fig. 5: Cross-dataset reconstructions for ICVL (left), Harvard (middle) and CAVE (right, downsampled to 256×256 to reveal blurring). **Top:** measurement, with the patch grid in red for ICVL and Harvard. **Middle:** reconstructed HSI projected to RGB, with markers at the sampled pixels; PSNR/SSIM are overlaid for ICVL and Harvard, and the RGB values of the metamer pair for CAVE. **Bottom:** per-pixel spectral radiance at the sampled pixels, with fill showing the predicted uncertainty (cf. Fig. 4).

ible with diffusion distillation techniques such as DMD [46], which can reduce sampling to a few steps, offering a path from minutes to seconds per scene.

Table 2 illustrates the effects of other design choices. In addition to toggling guidance, we evaluate smaller patch sizes, overlapping patches, and disabling patch rescaling ($c_{\text{lsq}}^p = 1$ in Eq. (3)). In the latter case, we trained a separate diffusion model without patch normalization (Sec. 3.2). Interestingly, reducing the patch size to 32 pixels—equal to the PSF kernel width—only marginally reduces performance, even though it makes the problem substantially more ill-posed. This shows that the cross-patch synchronization introduced by guidance plays a critical role in mitigating patch-based ambiguity, since removing it causes a larger PSNR drop for 32-pixel patches (33.27 to 29.27) than for 64-pixel patches (34.67 to 32.16). We also tested patches smaller than the PSF kernel size (e.g., 16×16 pixels), but these resulted in substantially worse reconstructions. Overlapping patches provide little benefit when guidance is active but become important otherwise, confirming that guidance corrects boundary artifacts. Finally, computing the patch scale factors c_{lsq}^p during sampling, instead of training the network to predict an exact per-patch scale, yields substantial improvements.

4.4 Comparison of PSF Designs and Model Interpretability

Table 3 shows how different PSF designs affect the reconstruction quality, using the eight PSFs depicted in Fig. 1b. For each lens, we render grayscale measurements using the ARAD1K dataset and train a separate diffusion model for

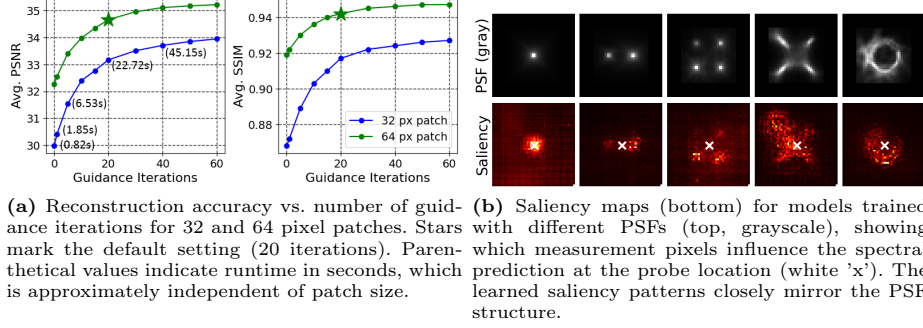


Fig. 6: Guidance iteration analysis (left) and input saliency visualization (right).

the same number of steps. Empirically, we find that reconstruction accuracy increases for lenses that induce stronger spatial-spectral mixing, but only to a certain extent. The T4 and R1 PSFs yield the best results, while the heavier mixing in the R2 and R3 designs from [24] cause a decline in performance, likely due to excessive blurring that diminishes spatial detail. These findings underscore the importance of balancing spatial and spectral encoding, and they suggest that the PSFs best suited for filterless hyperspectral imaging differ from those designed for RGB sensor mosaics [24].

To interpret what the patch-based diffusion models learn, we compute saliency maps via perturbation [39, 48]. For a probe pixel at location (r_x, r_y) in an output HSI patch, we define the saliency of each input measurement pixel (i, j) as $S(i, j \mid r_x, r_y) = \mathbb{E}_p [\sum_{\lambda} |\partial \mathbf{x}_0^p(r_x, r_y, \lambda) / \partial \mathbf{y}^p(i, j)|]$. We approximate this by systematically setting individual measurement pixels to zero, re-running the reconstruction for the patch (without guidance), and recording changes in the output spectrum at the probe location. We average the resulting saliency maps over 20 randomly sampled patches from the test set. As shown in Fig. 6b, the saliency map for each lens model closely resembles the PSF kernel used to generate the measurements, even though these kernels are not explicitly provided to the network during training. This suggests that our models implicitly learn the structure of the physical image formation process, prioritizing the pixels in the patch that optically map to each output location.

5 Experimental Prototype

5.1 Calibration and Setup

To validate our approach in a real imaging system, we construct the prototype shown in Fig. 7(a). We fabricate a 3 mm diameter metasurface composed of sub-wavelength TiO_2 nanocylinders following the R1 PSF design (Sec. 3.1); see Supp. Sec. S2 for fabrication details. The metasurface is mounted 4 cm in front of a monochrome CMOS sensor, with a 450–700 nm band-pass filter at the aperture.

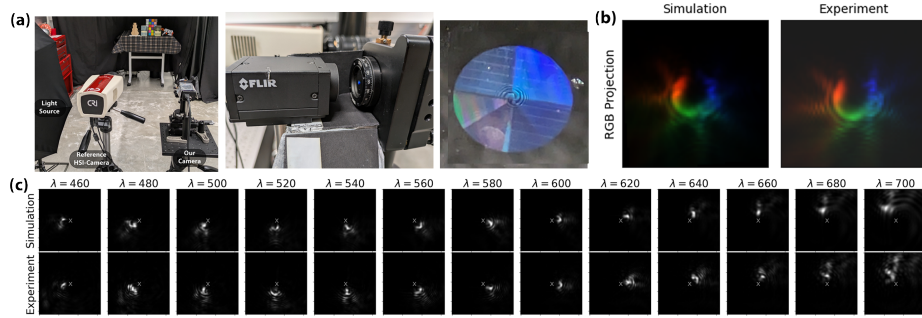


Fig. 7: (a) Imaging setup: a reference hyperspectral camera and our metalens prototype capture tabletop scenes side by side. Closeups show the prototype camera and the fabricated metasurface. (b,c) Comparison of the experimentally measured and simulated R1 PSFs, shown as RGB projections (top right) and per-wavelength slices from 460 to 700 nm.

Figure 7(b,c) compares the simulated PSF to the experimentally measured PSF captured using a tunable laser source, confirming excellent spatial agreement. Approximately 96% of the incident energy is concentrated within a 128×128 pixel ($832 \times 832 \mu\text{m}$) region at the sensor, which we use as the PSF kernel. Compared to the 64×64 kernel used in simulation (Sec. 4), this larger spatial support means that more light is scattered across patch boundaries when forming measurements, making the reconstruction more challenging. Separately, we calibrate the per-wavelength spectral scaling of the measured PSF using paired captures of planar calibration scenes; see Supp. Sec. S2.4 for details.

We pre-train our model from scratch using the experimentally measured PSF, and then we finetune it to bridge the gap between simulation and the physical system, where stray light extends beyond the truncated PSF kernel, and where the sensor’s spectral sensitivity and noise are not precisely known. To finetune, we need paired captures of real measurements and reference HSIs. For this, we place a commercial multi-shot non-compressive hyperspectral camera (Nuance FX, CRI Inc.) beside the prototype (Fig. 7a) and capture 15 indoor tabletop scenes with both cameras, crudely registering the two views using per-scene homographies (full details in the supplement). Because these scenes contain objects at varying depths, the homography yields only approximate spatial alignment. We adapt our model to tolerate this misalignment by introducing random spatial offsets during patch extraction in pre-training. Specifically, when pairing a measurement patch with its reference HSI patch, we apply a small random shift so that residual misalignment is modeled as an additional source of per-patch ambiguity in the learned conditional distribution. The resulting jitter in per-patch predictions is corrected by guidance at inference, which enforces consistency with the full-field measurement after stitching the patch predictions. This allows us to finetune on these imperfectly aligned pairs and still bridge the remaining sim-to-real gaps in spectral calibration, stray light, and sensor noise.

5.2 Training and Results

We pre-train the diffusion model on simulated HSI-measurement pairs rendered from the ARAD1K dataset using the scale-calibrated experimental PSF via Eq. (1). We assume the PSF is shift-invariant, which we validate experimentally in Supp. Sec. S2.3. Training uses random 64×64 patch crops from the full-field pairs. To prepare the model for finetuning using the imperfectly aligned supervision pairs described above, we introduce random spatial offsets of up to 12 pixels when extracting these crops. Gaussian noise is also added to simulated measurements; see Supp. Sec. S3.4 for details.

After pre-training, we freeze the model and finetune only the QKV attention matrices and the final projection head using rank-8 LoRA [14, 19], updating 166 K parameters (0.22% of the full model) with a learning rate of 1×10^{-6} for 10 epochs. We use six of the 15 tabletop scenes for finetuning and evaluate on the remaining eight. The finetuning and test scenes contain no shared objects. For each test scene, we reconstruct 26 spectral channels spanning 450–700 nm over a 640×960 pixel central crop (150 patches) of the full sensor measurement, corresponding to the region containing the scene objects. We use 10 guidance iterations to avoid over-guiding on noisy measurements; all other settings match Sec. 4. Reconstruction requires approximately 13 minutes per scene.

Figure 8 displays results for representative test scenes, with all scenes and full metrics in Supp. Sec. S4.3. Across the test set, our model achieves an average PSNR of 31.20 dB, SAM of 0.157 (8.9°), and SSIM of 0.894. These results confirm experimentally that HSIs can be reconstructed solely from lens-induced chromatic aberration in a filterless snapshot measurement.

Without guidance, the predicted spectra at each pixel varies substantially when repeating the reconstruction using different initial noise seeds (magenta fill in Fig. 8). Straight edges appear jagged and textured regions exhibit ghosting artifacts, as each patch prediction carries a random spatial offset. When guidance is applied, both effects are corrected by enforcing consistency with the full-field measurement: spatial artifacts are resolved and the spread of per-pixel spectral predictions becomes significantly narrower (green fill), often enveloping the true spectral radiance curve. As shown in Fig. 9, the predicted uncertainty is strongly correlated with reconstruction error. We sample 20K pixels from regions containing objects (indicated by the red contours) and compute a Pearson correlation coefficient of 0.69 between uncertainty and MSE, confirming experimentally that per-pixel uncertainty identifies regions where reconstruction is less trustworthy.

6 Limitations

Several limitations stem from the patch-based formulation. First, like all methods that encode spectral information into spatial structure via PSF blurring, our approach requires scenes with texture at the scale of the PSF support. A spatially homogeneous region produces a near-uniform measurement regardless

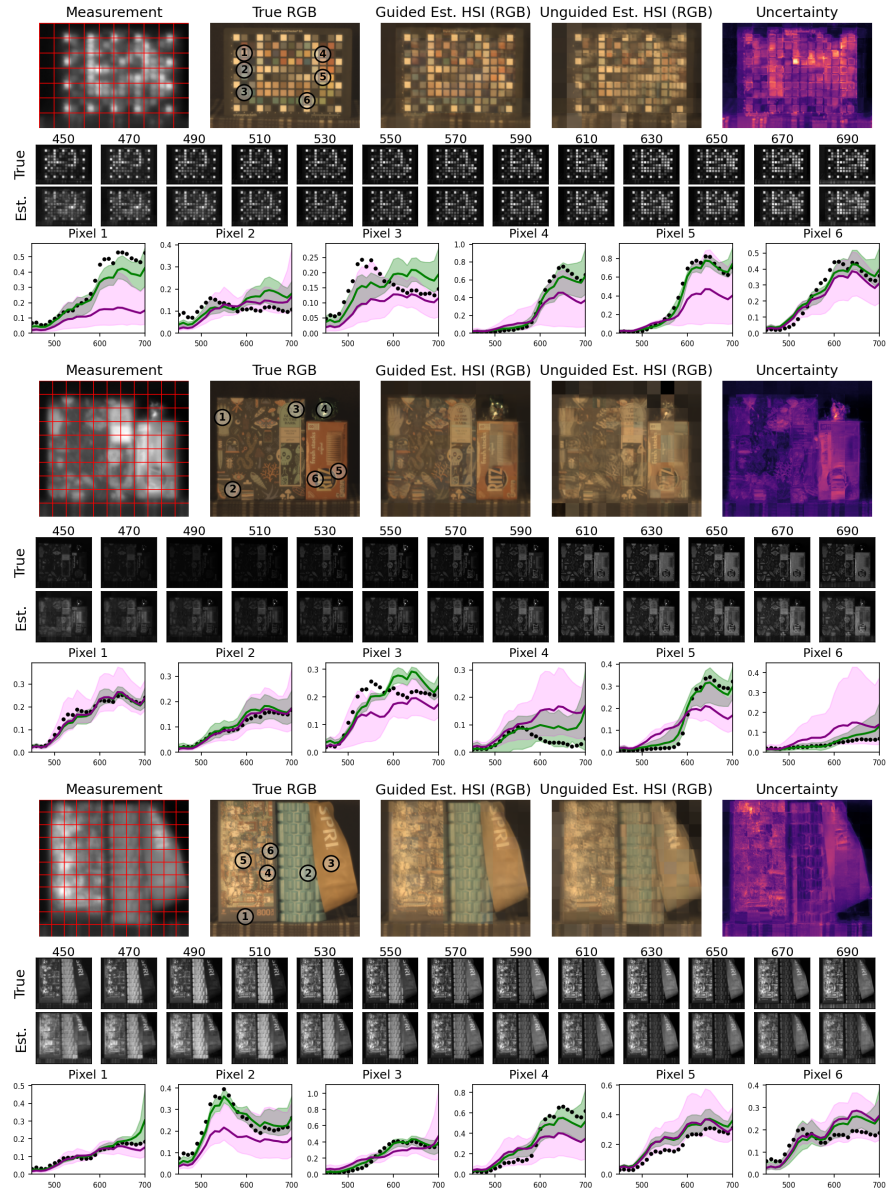


Fig. 8: Experimental reconstructions for three test scenes. Top rows: grayscale measurement (with patch grid in red), ground truth RGB, guided and unguided HSI estimates (projected to RGB), and per-pixel uncertainty. Middle rows: per-wavelength slices of the ground truth and estimated HSI. Bottom rows: spectral radiance at six marked pixels using guidance (green) and unguided (magenta) with uncertainty (fill)

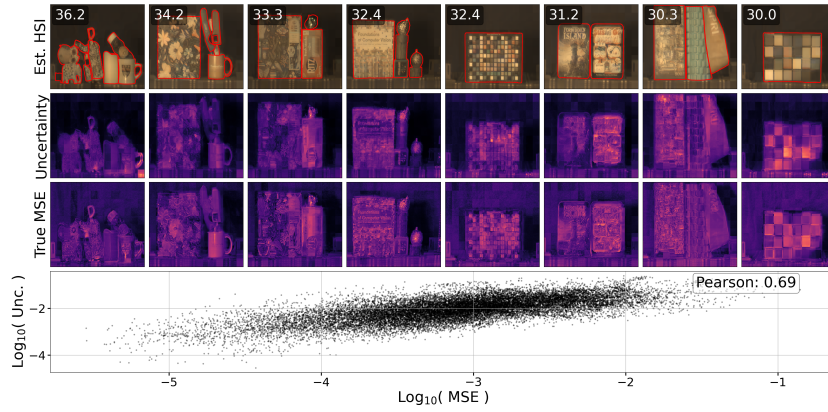


Fig. 9: Uncertainty versus reconstruction error on experimental test scenes (cf. Fig. 4). Top row: guided HSI estimates projected to RGB with PSNR overlaid.

of its spectrum, making the mapping from measurement to HSI many-to-one and the spectrum unrecoverable. The influence of spatial feature size on reconstruction accuracy is probed in a controlled synthetic experiment, where we find that reducing the spatial frequency of a checkerboard pattern increases reconstruction error (Supp. Sec. S4.7). This is also visible in the two ColorChecker test scenes (Fig. 9) where larger uniform squares are reconstructed notably worse than smaller ones. Second, because patches are reconstructed independently and guidance couples them only at shared edges, global context cannot propagate across the image, leading to uneven reconstruction quality within a scene (e.g., the red flowers in Fig. 5). An encoder that conditions each patch on a full-resolution image embedding could address this.

7 Conclusion

We present one of the first demonstrations that hyperspectral images can be reconstructed solely from the local chromatic aberration in a single grayscale measurement that is captured through a flat-optic lens. Central to this is our integration of patch diffusion models with guidance based on the camera’s point-spread function. By leveraging local diffusion while enforcing cross-patch consistency, this work provides a new approach for processing optically encoded measurements, as well as a new tradeoff between optical and computational complexity in snapshot hyperspectral imaging.

Acknowledgements

This work was supported in part by NSF cooperative agreement PHY2019786 (an NSF AI Institute, iaifi.org).

References

1. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: ECCV. pp. 19–34 (2016)
2. Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Van Gool, L., Liu, S., Li, Y., Feng, C., Lei, L., Li, J., Du, S., Wu, C., Leng, Y., Song, R., Zhang, M., Song, C., Zhao, S., Lang, Z., Wei, W., Zhang, L., Dian, R., Shan, T., Guo, A., Feng, C., Liu, J., Agarla, M., Bianco, S., Buzzelli, M., Celona, L., Schettini, R., He, J., Xiao, Y., Xiao, J., Yuan, Q., Li, J., Zhang, L., Kwon, T., Ryu, D., Bae, H., Yang, H.H., Chang, H.E., Huang, Z.K., Chen, W.T., Kuo, S.Y., Chen, J., Li, H., Liu, S., Sabarinathan, Uma, K., Bama, B.S., Roomi, S.M.M.: Ntire 2022 spectral recovery challenge and data set. In: CVPRW. pp. 863–881 (2022)
3. Arbabi, E., Arbabi, A., Kamali, S.M., Horie, Y., Faraon, A.: Multiwavelength metasurfaces through spatial multiplexing. *Scientific Reports* **6**, 32803 (2016)
4. Brookshire, C., Liu, Y., Chen, Y., Chen, W.T., Guo, Q.: Metahdr: Single shot high-dynamic range imaging and sensing using a multifunctional metasurface. *Opt. Express* **32**(15), 26690–26707 (2024)
5. Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., Gool, L.V.: Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In: CVPRW (2022)
6. Cai, Y., Lin, J., Wang, H., Yuan, X., Ding, H., Zhang, Y., Timofte, R., Van Gool, L.: Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In: *Advances in Neural Information Processing Systems* (2022)
7. Chakrabarti, A., Zickler, T.: Statistics of real-world hyperspectral images. In: CVPR. pp. 193–200 (2011)
8. Chen, N., Yue, J., Fang, L., Xia, S.: Spectraldiff: A generative framework for hyperspectral image classification with diffusion models. *IEEE Trans. Geoscience and Remote Sensing* **61**, 1–16 (2023)
9. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016)
10. Chung, H., Kim, J., McCann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: *ICLR* (2023)
11. Chung, H., Sim, B., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. In: *Advances in Neural Information Processing Systems* (2022)
12. Ding, K., Wang, M., Chen, M., Wang, X., Ni, K., Zhou, Q., Bai, B.: Snapshot spectral imaging: From spatial-spectral mapping to metasurface-based imaging. *Nanophotonics* **13**(8), 1303–1330 (2024)
13. Fu, Q., Souza, M., Choi, E., Shin, S., Baek, S.H., Heidrich, W.: Limitations of data-driven spectral reconstruction: An optics-aware analysis. *IEEE Transactions on Computational Imaging* (2025)
14. Guo, H., Li, Y., Dai, T., Xia, S.T., Benini, L.: Intlora: Integral low-rank adaptation of quantized diffusion models. In: *Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 267, pp. 20858–20879. PMLR (2025)
15. Han, X., Zickler, T., Nishino, K.: Multistable shape from shading emerges from patch diffusion. *NeurIPS* (2024)
16. Hazineh, D., Lim, S.W.D., Guo, Q., Capasso, F., Zickler, T.: Polarization multi-image synthesis with birefringent metasurfaces. In: *ICCP*. pp. 1–12 (2023)

17. Hazineh, D.S., Lim, S.W.D., Shi, Z., Capasso, F., Zickler, T., Guo, Q.: D-flat: A differentiable flat-optics framework for end-to-end metasurface visual sensor design. arXiv preprint arXiv:2207.14780 (2022)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (2020)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
20. Hu, J., Song, B., Fessler, J.A., Shen, L.: Patch-based diffusion models beat whole-image models for mismatched distribution inverse problems. arXiv preprint arXiv:2410.11730 (2024)
21. Hu, J., Song, B., Xu, X., Shen, L., Fessler, J.A.: Learning image priors through patch-based diffusion models for solving inverse problems. In: Advances in Neural Information Processing Systems (2024)
22. Hu, X., Cai, Y., Lin, J., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In: CVPR. pp. 17542–17551 (2022)
23. Huang, T., Dong, W., Yuan, X., Wu, J., Shi, G.: Deep gaussian scale mixture prior for spectral compressive imaging. In: CVPR. pp. 16216–16225 (2021)
24. Jeon, D.S., Baek, S.H., Yi, S., Fu, Q., Dun, X., Heidrich, W., Kim, M.H.: Compact snapshot hyperspectral imaging with diffracted rotation. ACM TOG **38**(4), 117:1–13 (2019)
25. Khorasaninejad, M., Zhu, A.Y., Roques-Carmes, C., Chen, W.T., Oh, J., Mishra, I., Devlin, R.C., Capasso, F.: Polarization-insensitive metalenses at visible wavelengths. Nano Letters **16**(11), 7229–7234 (2016)
26. Khorasaninejad, M., Capasso, F.: Metalenses: Versatile multifunctional photonic components. Science **358**(6367), eaam8100 (2017)
27. Li, K., Dai, D., Van Gool, L.: Jointly learning band selection and filter array design for hyperspectral imaging. In: WACV. pp. 6373–6383 (2023)
28. Li, M., Fu, Y., Zhang, Y.: Spatial-spectral transformer for hyperspectral image denoising. In: AAAI (2023)
29. Liu, W., Zhang, C., Li, J., Zhang, H., Pang, M., Ma, S., Liu, Y., Wang, Y., Song, W.: Grayscale-driven snapshot hyperspectral imaging via efficient phase modulation. Optics Communications **596**, 132459 (2025). <https://doi.org/https://doi.org/10.1016/j.optcom.2025.132459>, <https://www.sciencedirect.com/science/article/pii/S0030401825009873>
30. Lv, T., Ye, H., Yuan, Q., Shi, Z., Wang, Y., Wang, S., Cao, X.: Aperture diffraction for compact snapshot spectral imaging. In: ICCV. pp. 10574–10584 (October 2023)
31. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: ECCV (2020)
32. Monakhova, K., Yanny, K., Aggarwal, N., Waller, L.: Spectral diffusercam: Lensless snapshot hyperspectral imaging with a spectral filter array. Optica **7**(10), 1298–1307 (2020)
33. Nie, S., Gu, L., Zheng, Y., Lam, A., Ono, N., Sato, I.: Deeply learned filter response functions for hyperspectral reconstruction. In: ICCV. pp. 4767–4776 (2018)
34. Pan, Z., Zeng, H., Cao, J., Zhang, K., Chen, Y.: Diffsci: Zero-shot snapshot compressive imaging via iterative spectral diffusion model. In: CVPR. pp. 25297–25306 (2024)
35. Pang, L., Rui, X., Cui, L., Wang, H., Meng, D., Cao, X.: Hir-diff: Unsupervised hyperspectral image restoration via improved diffusion models. In: CVPR. pp. 3005–3014 (2024)

36. Romero, J., Fu, Q., Ravasi, M., Heidrich, W.: Uncertainty quantification in hsi reconstruction using physics-aware diffusion priors and optics-encoded measurements (2025), <https://arxiv.org/abs/2511.18473>
37. Salesin, K., Seyb, D., Friday, S., Jarosz, W.: Diy hyperspectral imaging via polarization-induced spectral filters. In: ICCP. pp. 1–12 (2022)
38. Sigger, N., Vien, Q.T., Nguyen, S.V., et al.: Unveiling the potential of diffusion model-based framework with transformer for hyperspectral image classification. *Scientific Reports* **14**, 8438 (2024)
39. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
41. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. *Appl. Opt.* **47**(10), B44–B51 (2008)
42. Wang, X., Wang, L., Ma, X., Zhang, M., Zhu, L., Huang, H.: In2set: Intra-inter similarity exploiting transformer for dual-camera compressive hyperspectral imaging. In: CVPR. pp. 24881–24891 (2024)
43. Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., Zhou, M.: Patch diffusion: Faster and more data-efficient training of diffusion models. In: *Advances in Neural Information Processing Systems* (2023)
44. Yao, Z., Liu, S., Yuan, X., Fang, L.: Specat: Spatial-spectral cumulative-attention transformer for high-resolution hyperspectral image reconstruction. In: CVPR. pp. 25368–25377 (2024)
45. Yasuma, F., Mitsunaga, T., Iso, D., Nayar, S.K.: Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE TIP* **19**(9), 2241–2253 (2010)
46. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: CVPR (2024)
47. Yuhas, R.H., Goetz, A.F.H., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In: *Summaries of the Third Annual JPL Airborne Geoscience Workshop*. vol. 1. JPL (1992)
48. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833 (2014)
49. Zeng, H., Cao, J., Zhang, K., Chen, Y., Luong, H., Philips, W.: Unmixing diffusion for self-supervised hyperspectral image denoising. In: CVPR. pp. 27820–27830 (2024)
50. Zhang, L., Luo, X., Li, S., Shi, X.: R2h-cdd: Hyperspectral imagery generation from rgb images based on conditional cascade diffusion probabilistic models. In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. pp. 7392–7395 (2023)
51. Zhang, Q., Yu, Z., Liu, X., Wang, C., Zheng, Z.: End-to-end joint optimization of metasurface and image processing for compact snapshot hyperspectral imaging. *Optics Communications* **530**, 129154 (2023)
52. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE TPAMI* **45**(8), 10346–10357 (2023)